UFPE-Essex-UNED at ImageCLEFmed Caption 2026: Training-Free Multimodal RAG for Medical Image Captioning

Heitor Pereira[1], Pedro Maranhão[1], Vitória Xavier[1], Tsang Ing Ren[1], Alba García Seco de Herrera[2] and Vahid Abolghasemi[3]

[1]*Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Recife, Brazil*
[2]*National University of Distance Education - UNED, Spain*
[3]*School of Computer Science & Electronic Engineering, University of Essex, CO4 3SQ, Colchester, U.K.*

Abstract

This paper presents the UFPE–Essex–UNED team's participation in the ImageCLEFmed Caption 2026 challenge. The team participated in the Caption Prediction and Caption Prediction Synthetical subtasks, employing a fully training-free pipeline based on retrieval-augmented generation (RAG) and prompt engineering, without any model fine-tuning. The approach used Llama 4 Maverick as the caption-generation model and MedSigLIP to construct the image-vector database for retrieval. In the Caption Prediction subtask, the approach achieved 2nd place with an overall score of 0.3603 and a Similarity of 0.99. For the Caption Prediction Synthetical subtask, the same pipeline, with a simplified configuration, achieved 4th place in overall score (0.4625) and 1st place in Similarity (0.98). The experimental results highlight the effectiveness of retrieval augmentation and prompt alignment for medical image captioning, even in fully training-free settings.

Keywords

Vision Language Model, Retrieval Augmented Generation, Large Language Models, Medical Image Captioning

1. Introduction

Medical image captioning has become an increasingly important research topic in multimodal Artificial Intelligence (AI), with potential applications in clinical documentation, medical education, image retrieval, and decision support systems. Unlike general-domain image captioning, generating captions for medical images requires clinically meaningful, semantically accurate descriptions, since even small linguistic inaccuracies can alter the interpretation of medical findings.

To support the development and evaluation of automatic medical image understanding systems, the ImageCLEFmed Caption task [1], organized within the ImageCLEF lab [2] of the Conference and Labs of the Evaluation Forum (CLEF), has established itself as one of the main benchmarking challenges in the field. Over the years, the task has evolved, incorporating progressively larger, curated datasets and evaluation methodologies focused on both semantic relevance and factual correctness. In this context, the adoption of Image-Caption Similarity [3] as a metric in 2023, further emphasized the importance of measuring the semantic alignment through reference-free metrics.

Recent advances in large multimodal foundation models and retrieval-augmented generation (RAG) [4] systems have created new opportunities for medical image captioning without requiring task-specific training. In particular, integrating large language models (LLMs) with retrieval mechanisms enables the use of external medical knowledge and example-based prompting to improve caption quality and contextual consistency. Training-free approaches that combine retrieval mechanisms with prompt-based

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
hmp@cin.ufpe.br (H. Pereira); pdm@cin.ufpe.br (P. Maranhão); vax@cin.ufpe.br (V. Xavier); tir@cin.ufpe.br (T. I. Ren); alba.garcia@lsi.uned.es (A. García Seco de Herrera); v.abolghasemi@essex.ac.uk (V. Abolghasemi)
https://www.uned.es/universidad/docentes/informatica/alba-garcia-seco-de-herrera.html/ (A. García Seco de Herrera); https://www.essex.ac.uk/people/ABOLG46201/Vahid-Abolghasemi (V. Abolghasemi)
0009-0008-3615-3849 (H. Pereira); 0000-0003-3082-1523 (P. Maranhão); 0000-0001-5731-1908 (V. Xavier); 0000-0002-3677-0264 (T. I. Ren); 0000-0002-6509-5325 (A. García Seco de Herrera); 0000-0002-2151-5180 (V. Abolghasemi)

generation offer a promising alternative to fine-tuning; however, their effectiveness in medical image captioning remains underexplored.

Previous work from our group [5] evaluated several state-of-the-art generative AI models for medical image captioning on the Radiology Objects in Context Version 2 (ROCOv2) dataset [6], including GPT-4o [7], Gemini 2.5 Pro [8] and Llama 4 [9]. The study analyzed both zero-shot and few-shot prompting strategies using multiple automated captioning metrics, highlighting the potential of large multimodal models for understanding radiology images.

In this work, we present the submission of the UFPE–Essex–UNED collaboration to the ImageCLEF 2026 Caption Prediction task. The approach employs a fully training-free pipeline based on RAG, combining Llama 4 Maverick [9] for caption generation, few-shot prompting, and multimodal retrieval using CLIP [10] and MedSigLIP [11] embeddings. Using this approach, the system achieved 2nd place in the competition. The main contribution of this paper is a fully training-free multimodal RAG pipeline that combines domain-specialized retrieval (MedSigLIP) with prompt alignment to dataset-specific caption styles, demonstrating that competitive performance can be achieved without fine-tuning large models.

2. Evaluation protocol

All experiments were conducted following the official evaluation protocol defined by the ImageCLEF 2026 Caption Prediction task organizers. This includes using predefined training, validation, and test splits. For the test data, reference annotations are not publicly available, and all submissions are evaluated centrally by the organizers. A detailed description of the evaluation metrics, their computation, and the ranking methodology is provided in the official overview paper of the ImageCLEF 2026 campaign [1].

The evaluation in the ImageCLEF 2026 Caption Prediction task is based on a combination of complementary metrics designed to capture different aspects of caption quality. The final system ranking is determined by an overall score that aggregates all evaluation dimensions, including BERTScore [12], ROUGE [13], Similarity [14] [3], BLEURT [15], MedCAT [16] [17], and AlignScore [18]. Although each metric captures a specific dimension of caption quality, their combination into a single overall score may obscure important differences between systems. Therefore, individual metrics remain essential to better understand system behavior and performance across linguistic quality, semantic alignment, and clinical correctness. This paper places particular emphasis on the Similarity component, a metric based on multimodal representations that measures the semantic alignment between images and generated captions. This metric was introduced in the 2023 challenge following the reference-free evaluation logic proposed by CLIPScore in 2021 [3], and it enables reference-free assessment of image–text correspondence. Notably, our systems achieved particularly strong results according to this metric.

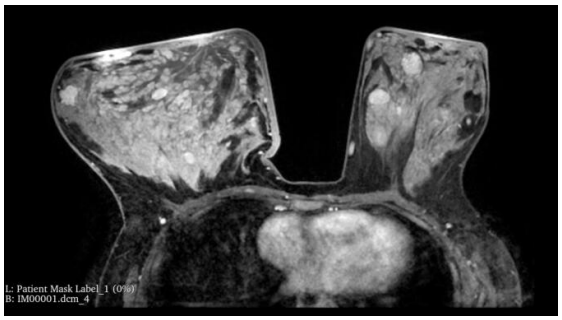
2.1. Dataset

The experiments were conducted using the dataset provided by the ImageCLEF 2026 Caption Prediction challenge organizers [2]. The development data is based on an updated and extended version of the ROCOv2 [6]. As in previous editions of the challenge, the images originate from biomedical articles contained in the PubMed Central (PMC) OpenAccess subset.

The dataset consists of curated radiology images paired with associated captions and manually controlled UMLS concept annotations provided as metadata. The captions resemble figure descriptions commonly found in biomedical literature, focusing primarily on descriptive and visually grounded content rather than exhaustive clinical reporting, as illustrated in Figure 2.

For the Caption Prediction subtask, the organizers provided 97364 training images, 19240 validation images, and 15249 test images, with the test split composed of previously unseen samples. The dataset distributed for the challenge extends the original ROCOv2 release described in the literature by incorporating additional and more recent PMC-derived radiology images, similarly to the dataset expansion performed in the ImageCLEFmedical Caption 2025 edition [21].

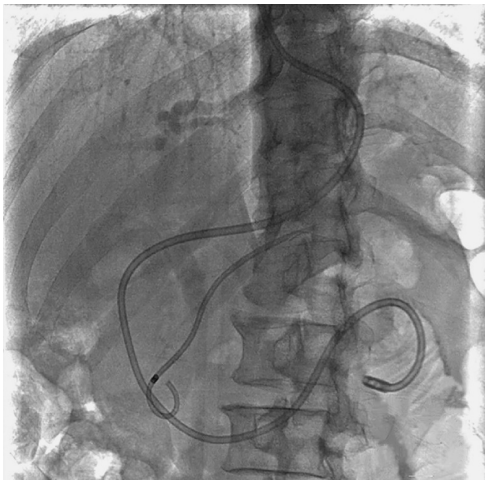
ImageCLEFmedical_Caption_2026_valid_2405



(a) **Ground Truth Caption**: Preoperative radiologic findings: magnetic resonance imaging findings. Severely dense fibroglandular tissue, diffuse parenchymal inflammation with skin thickening, and areolar edema in the right breast.

(b) **Predicted Caption**: Axial T1-weighted fat-saturated post-contrast MRI of the breasts showing multiple enhancing lesions in both breasts.

ImageCLEFmedical_Caption_2026_valid_2415



(c) **Ground Truth Caption**: Successful placement of a stent.

(d) **Predicted Caption**: Fluoroscopic image showing a catheter navigating through the inferior vena cava and into the hepatic vein.

Figure 1: Examples of biomedical figures and captions extracted from the 2026 ImageCLEFcaption dataset. The first image (Pakbaz et al., CC BY [19]) provides a more detailed and self-contained description while the second (Liang et al., CC BY [20]) depends heavily on the surrounding article context.

Additionally, the captions distributed by the organizers were preprocessed by removing hyperlinks and formatting artifacts. This edition of the challenge also introduced a new Caption Prediction Synthetical subtask, in which captions were automatically generated from the images. For this task, the organizers provided 97222 training images, 19239 validation images, and 15262 test images.

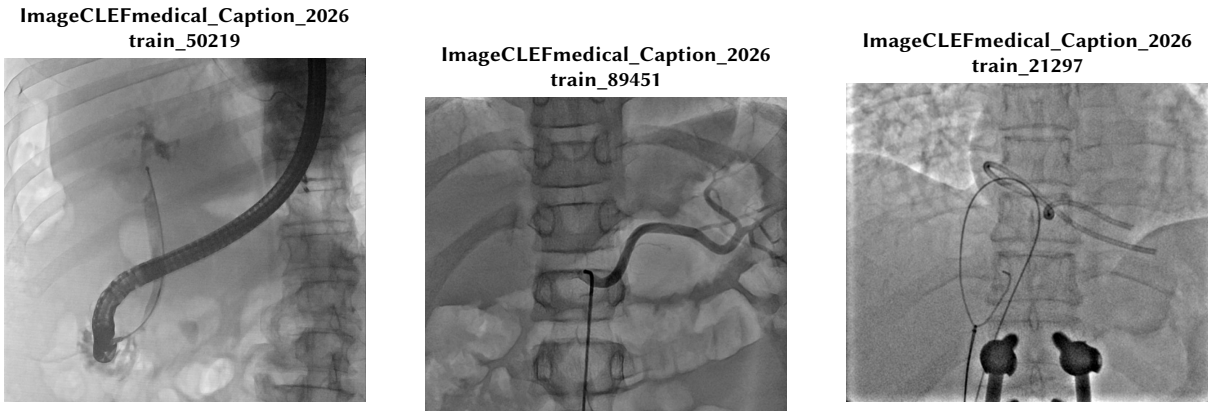
2.2. Evaluation Metrics

The evaluation protocol for the Caption Prediction task follows the official CLEF Caption Evaluation benchmark. Participant ranking is determined by the average score across six complementary metrics assessing both relevance and factual consistency.

Relevance is evaluated using Image–Caption Similarity [14] [3], BERTScore Recall [12] with inverse document frequency (idf) weighting, ROUGE-1 F-measure [13], and BLEURT [15]. According to the official evaluation implementation [22], the Image–Caption Similarity metric measures the semantic similarity between embeddings extracted from the medical image and the generated caption using the MedImageInsight model [14], following the idea from CLIPScore [3]. The Similarity score reflects the degree of alignment between the visual content and the generated textual description in a shared multimodal embedding space.

BERTScore computes contextual semantic similarity between generated and reference captions using contextualized embeddings obtained from the `microsoft/deberta-xlarge-mnli` model [23]. In the official evaluation, the recall variant with inverse document frequency weighting is employed to emphasize semantically informative tokens. ROUGE-1 evaluates lexical overlap by measuring unigram-level agreement between generated and reference captions, while BLEURT employs a learned BERT-based evaluation model trained to approximate human judgments of text quality. Prior to computing text-based relevance metrics, captions are normalized by converting all text to lowercase, replacing numbers with a generic token, and removing punctuation.

Factual consistency is assessed through UMLS Concept F1 and AlignScore. UMLS Concept F1 extracts clinical entities (UMLS concepts) from generated and reference captions using MedCAT [16]. The F1-



(a) **Caption**: Cholangiogram demonstrating filling defects at the common hepatic duct.

(b) **Caption**: Angiogram showing the splenic artery.

(c) **Caption**: Showing wire and snare loop being pulled down to reposition the catheter.

Figure 2: Retrieved image-caption pairs used as examples for captioning image: ImageCLEFmedical_Caption_2026_valid_2415 1d (RAG Model: MedSigLIP). Image attributions (left to right): Cheung et al., CC BY-NC [26]; Jawale et al., CC BY [27]; Shah et al., CC BY [28].

score is then computed over the identified UMLS concepts, restricted to the clinically relevant semantic types defined by the MEDCON evaluation protocol [17]. AlignScore [18] evaluates factual alignment using a RoBERTa-based [24] model that measures whether the claims present in the generated caption are supported by the reference text. The final challenge score is obtained by averaging all metric scores.

3. Methodology

This section describes the proposed pipeline used in our ImageCLEF 2026 submission. It presents the overall workflow for caption generation, including the retrieval process, prompt engineering strategy, and the multimodal models employed throughout the experiments. We also describe the different configurations evaluated during the competition and the progressive modifications introduced to improve system performance.

3.1. Pipeline Outline

The methodology was motivated by a previous work [5], in which we conducted a comparative evaluation of state-of-the-art generative AI models for medical image captioning on the ROCOv2 dataset. In that study, models including GPT-4o, Gemini 2.5 Pro, Claude Sonnet, and Llama 4 were evaluated under zero-shot and few-shot prompting settings using standard captioning metrics such as BLEU, ROUGE, METEOR [25], and BERTScore. The experiments demonstrated the strong performance of Llama 4 Maverick for radiology caption generation, thereby motivating its adoption as the caption-generation model in the present work. Additionally, the pilot experiments in the previous study provided guidance on the (k=3) of retrieved examples to incorporate into the RAG pipeline.

The approach is based on a training-free multimodal RAG pipeline combined with few-shot prompting. No additional model training or fine-tuning was performed during the competition. Instead, the method relies on retrieving visually similar medical images from the training set and leveraging their associated captions as contextual examples for the language model.

For each image in the training set, visual embeddings were extracted using a vision-language encoder and stored in a vector database. Then, during inference, as illustrated in Figure 3, the embedding of the query image was compared with the stored embeddings using cosine similarity, and the three most similar images were retrieved. The captions associated with these retrieved images were then incorporated into the prompt as few-shot examples. Finally, the domain-tuned prompt, together with the query image, was provided to the large language model to generate the final caption.

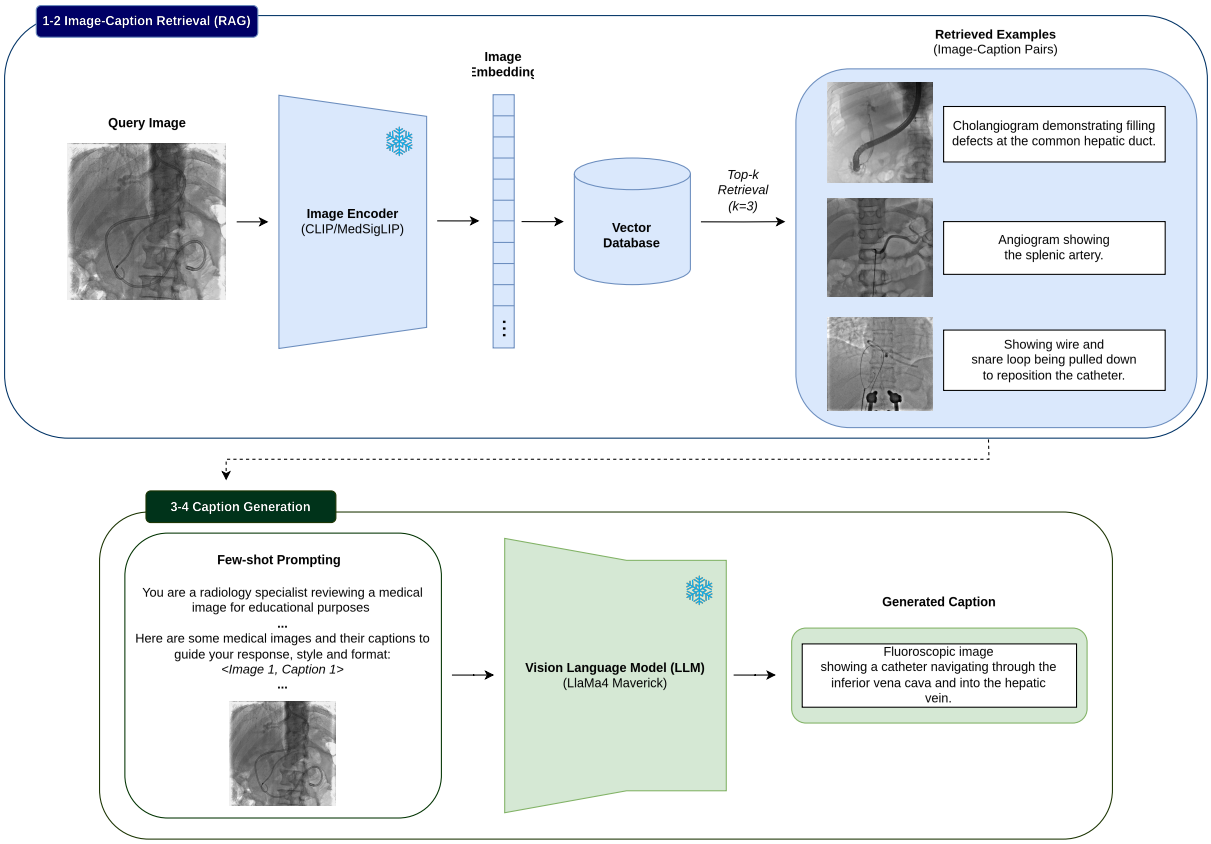


Figure 3: Overview of the proposed training-free multimodal RAG pipeline.

The inference pipeline can be summarized as follows:

1. Extract the embedding of the query image.
2. Retrieve the three most similar images from the vector database.
3. Incorporate the retrieved image-caption pairs into the prompt as few-shot examples.
4. Generate the final caption using the vision language model.

Different embedding models and prompt configurations were evaluated throughout the competition to improve retrieval quality and caption generation performance. Initially, to accelerate experimentation, we constructed the vector database using only 10k samples from the training set and employed CLIP ViT-B/32 [10] as the retrieval encoder. This initial configuration allowed for rapid iteration and validation of the proposed RAG pipeline.

Subsequently, we refined the prompting strategy and expanded the vector database to include the entire training dataset. Increasing the retrieval pool improved the diversity and relevance of the retrieved examples used during the few-shot prompting stage. Finally, the retrieval encoder was replaced with MedSigLIP, the vision encoder introduced in MedGemma [11]. MedSigLIP is a medically adapted version of SigLIP [29], optimized for medical image-text alignment, enabling the retrieval system to retrieve semantically and clinically more relevant nearest neighbors for caption generation.

3.2. Model

The model `meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8`, accessed through the DeepInfra API, was used to conduct all experiments. Model interactions were implemented using LangChain to provide a unified interface for LLM inference and facilitate reproducibility [5]. The inference settings used throughout the experiments are summarized in Table 1. A sufficiently large maximum output token limit was specified to avoid premature truncation during generation, while all remaining decoding parameters, including top-p, were left at their default values.

Table 1
Inference settings used for the caption generation model throughout all experiments.

Parameter	Value
Model	`meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8`
Deployment	DeepInfra API
Framework	LangChain
Temperature	0.1
Maximum output tokens	1,000,000
Top-p	Default
Other decoding parameters	Default

Additional implementation details, including the complete inference pipeline, prompts, and source code, are available in our public GitHub repository.[1]

3.3. Prompt Engineering

We designed and iteratively refined the prompting strategy to align the generated captions with the stylistic and semantic characteristics of the ImageCLEF dataset. The prompting process evolved from a general radiology-oriented formulation toward a dataset-specific design emphasizing concise, descriptive, and visually grounded figure captions.

Initial Prompt

```
You are a radiology specialist reviewing a medical image
for educational purposes. Your job is to provide a detailed
and accurate caption for the medical image.

Use these instructions to guide your response:
- Your caption should be assertive, concise,
  and clinically accurate
- Focus on describing the visible anatomical structures,
  clinical findings, and medical devices
- Include any abnormalities, diagnostic findings,
  or interventions visible in the image
- Use appropriate medical terminology
- Be specific about the imaging modality,
  body region, and orientation when applicable

Following these steps:
Step 1: Analyze the image carefully and identify all
visible anatomical structures, clinical findings,
and medical devices

Step 2: Note any abnormalities, diagnostic findings,
or interventions visible in the image

Step 3: Generate a clear and informative caption
that describes the image comprehensively

Provide your response as a JSON object:

{
    "caption": "A detailed and clinically accurate
    description of the medical image"
}
```

The initial prompt framed the model as a radiology specialist focused on generating clinically accurate

[1] https://github.com/PedroDidier/experiments-clef

descriptions of medical images.

While this prompt encourages clinically detailed descriptions, it may lead the model to produce overly interpretative or diagnostic statements that are not aligned with the dataset annotations. Since the ImageCLEF captions resemble figure descriptions from biomedical articles rather than full clinical reports, we refined the prompt to better match this writing style.

Refined Prompt

```
You are a radiology specialist reviewing a medical image
for educational purposes. Your job is to provide a detailed
and accurate caption for the medical image.

The target style is based on captions from biomedical
literature (PMC Open Access / ROCOv2 dataset), where
captions are written as figure descriptions accompanying
journal articles.

These captions are descriptive, concise, and focus on
visible content rather than full diagnostic interpretation.

Use these instructions to guide your response:
- Your caption should be assertive, concise,
  and clinically accurate
- Focus on describing visible anatomical structures,
  imaging findings, and medical devices
- Include abnormalities, interventions, or notable
  findings only when clearly visible
- Do not make assumptions or infer findings that
  are not visually supported
- Avoid speculative diagnoses
- Prioritize objective image description,
  similar to biomedical article figure captions

Following these steps:
Step 1: Analyze the image carefully and identify all
visible anatomical structures, findings, and medical devices

Step 2: Note any clearly visible abnormalities,
interventions, or pathologies

Step 3: Generate a concise and informative caption
consistent with biomedical article figure captions

Provide your response as a JSON object:

{
    "caption": "A detailed and clinically accurate
    description of the medical image"
}
```

This refinement reduces speculative language and enforces a more controlled and descriptive generation style, improving alignment with the dataset characteristics and the evaluation metrics, particularly those based on semantic similarity and lexical overlap.

We incorporated the retrieved images and their corresponding captions into the prompt as contextual examples using a few-shot prompting strategy. The following snippet was appended to the end of the prompt for both prompt variants described previously, while the remainder of the prompt remained unchanged.

RAG Addition

```
... <previous domain specific prompt>
Here are some similar medical images and their captions
as examples to guide your response style and format
Example 1: {
  "image": <encoded image>
  "caption": "Pretreatment chest X-ray"
}
```

By incorporating retrieved image-caption pairs as few-shot examples, the model is exposed to visually and semantically similar cases. This mechanism improves semantic grounding and guides the model toward generating captions that are consistent in structure, terminology, and level of detail with the dataset.

As shown in Section 4, the combination of prompt refinement and retrieval-based few-shot conditioning leads to consistent improvements across evaluation metrics.

4. Results

This section presents the experimental results obtained for both Caption Prediction subtasks and analyze the impact of different retrieval and prompting configurations within the proposed RAG pipeline.

4.1. Caption Prediction

Table 2 summarizes the configurations of the submitted runs to the Caption Prediction task. The variations include the use of a reduced versus full retrieval database, the transition from an initial to a refined prompt design, and the comparison between CLIP and MedSigLIP as embedding models. These configurations allow the analysis of how retrieval quality and prompt engineering influence overall system behavior and performance.

Table 2
Comparison of retrieval database size, prompting strategy, and embedding model across submitted runs to the ImageCLEF 2026 Caption Prediction task.

Run	Retrieval Database	Prompting Strategy	Embedding Model
Run 1	10k subset	Initial prompt	CLIP
Run 2	Full dataset	Refined prompt	CLIP
Run 3	Full dataset	Refined prompt	MedSigLIP

Table 3 presents the detailed performance of our submitted runs compared to the best-performing system in the ImageCLEF 2026 Caption Prediction subtask. The results show that the initial configuration (Run 1) already achieved competitive performance, obtaining an overall score of 0.3433 and ranking 15th out of 54 submissions. Despite relying on a reduced vector database containing only 10k training samples and using CLIP ViT-B/32 as the retrieval encoder, this configuration would have placed the system among the top 6 submissions in the leaderboard. This suggests that the proposed training-free pipeline is effective even under constrained retrieval settings.

The largest performance improvement was observed after expanding the vector database to the full training set and refining the prompting strategy. This modification increased the overall score from 0.3433 in Run 1 to 0.3554 in Run 2, improving the ranking to 9th out of 54 submissions and a 4th place in the leaderboard. This result demonstrates the importance of retrieval diversity and effective prompt alignment with the dataset characteristics. Finally, replacing the retrieval encoder with MedSigLIP yielded an additional improvement, increasing the overall score to 0.3603 in Run 3 and resulting in a 2nd place ranking in the Caption Prediction task leaderboard. Although the gain introduced by MedSigLIP was smaller compared to the previous modification, it consistently improved most evaluation metrics.

Table 3
Evaluation results on the test set across multiple metrics for the submitted runs in the ImageCLEF 2026 Caption Prediction subtask, compared with the best-performing system of the challenge. Bold values indicate the best-performing result among the submitted runs of our team for each evaluation metric.

Run	Overall	Relevance	Factuality	BERT	ROUGE	Similarity	BLEURT	MedCAT	Align
Run 1	0.3433	0.5408	0.1459	0.5903	0.2527	0.9763	0.3439	0.1606	0.1311
Run 2	0.3554	0.5468	0.1641	0.5976	0.2668	0.9784	0.3443	0.1733	**0.1549**
Run 3	**0.3603**	**0.5527**	**0.1679**	**0.6017**	**0.2725**	**0.9902**	**0.3463**	**0.1822**	0.1537
Best ImageCLEF run	0.3755	0.5786	0.1724	0.6250	0.3234	1.0104	0.3555	0.2160	0.1287

Among the configurations, Run 3 achieves the strongest overall performance across most evaluation metrics. In particular, it attains the highest scores among the runs in BERTScore (0.6017), ROUGE (0.2725), Similarity (0.9902), BLEURT (0.3463), and MedCAT (0.1822), indicating consistent improvements in both linguistic quality and domain-specific correctness when using the refined prompt and MedSigLIP embeddings. Although Run 2 slightly outperforms Run 3 in the Alignment metric, the difference is marginal, suggesting a minor trade-off between alignment and other quality aspects.

When compared to the top-ranked ImageCLEF system, the best configuration remains competitive, particularly in multimodal similarity, where it achieves a Similarity of 0.9902 versus 1.0104. It is worth noting that the Similarity implementation includes a scaling factor that can result in values above 1.

4.2. Caption Prediction Synthetical

Table 4 presents the performance of our submitted run in the ImageCLEF 2026 Caption Prediction Synthetical subtask. For this subtask, the system achieved an overall score of 0.4625, placing 4th among the participating submissions. Although the system ranked below the top three in the overall metric, it achieved the highest Similarity among all participating teams, with a value of 0.9843. This result indicates a strong semantic alignment between the generated captions and the corresponding medical images. Interestingly, the absolute scores obtained in the synthetic caption task were higher than those observed in the standard caption task, including values exceeding the first-place submission of the standard track. This difference suggests that the synthetic captions may exhibit linguistic or structural characteristics that are easier for large multimodal language models to reproduce consistently.

The experimental configuration adopted for the synthetic caption task corresponds to the same setup used in Run 1 of the standard caption experiments described in Table 2. In contrast to the standard caption task, we did not perform additional iterations involving prompt refinement, expansion of the retrieval database, or modifications to the retrieval encoder. During the competition, most development efforts were concentrated on the standard caption track due to the greater availability of intermediate leaderboard feedback, and therefore, the synthetic caption experiments were conducted using a simpler version of the proposed pipeline. Despite this simplified configuration, the system remained highly competitive, particularly in terms of multimodal similarity.

Table 4
Evaluation results on the test set across multiple metrics for the submitted run in the ImageCLEF 2026 Caption Prediction Synthetical subtask, compared with the best-performing system of the challenge. Bold values indicate the best-performing result among the submitted runs of our team for each evaluation metric.

Configuration	Overall	Relevance	Factuality	BERT	ROUGE	Similarity	BLEURT	MedCAT	Align
Run 1	**0.4625**	**0.6372**	**0.2878**	**0.6621**	**0.4795**	**0.9843**	**0.4230**	**0.3782**	**0.1974**
Best ImageCLEF run	0.5840	0.7183	0.4496	0.7580	0.6353	0.9737	0.5063	0.5332	0.3660

5. Discussion

5.1. Caption Prediction

The experimental results highlight the effectiveness of RAG and prompt engineering for medical image captioning, even without any task-specific fine-tuning. The strong performance achieved by the initial configuration suggests that large multimodal language models already possess substantial transferable knowledge for radiology image understanding.

The refinement of the prompt was particularly effective: since the captions in the challenge dataset correspond directly to figure descriptions extracted from biomedical articles, aligning the prompt with this descriptive style proved more effective than encouraging detailed diagnostic reporting. This indicates that prompt alignment with dataset characteristics can substantially influence caption quality.

Similarly, using a medically specialized retrieval encoder suggests that domain-specific retrieval representations can enhance retrieval quality and provide more clinically relevant contextual examples for the language model.

The retrieval-augmented pipeline proved particularly effective in capturing image–text semantic alignment. However, the best-performing system achieves higher scores in text-based and domain-specific metrics such as ROUGE, BLEURT, and MedCAT, suggesting stronger performance in lexical overlap and factual consistency. These findings indicate that while the training-free approach excels in semantic alignment, there remains room for improvement in capturing fine-grained clinical and linguistic details. Overall, the experiments demonstrate that competitive medical image captioning systems can be constructed using training-free pipelines that combine large language models, RAG, and carefully designed prompts.

5.2. Caption Prediction Synthetical

A notable observation from the synthetic task concerns the strong multimodal similarity of the generated captions. According to the official evaluation implementation [22], this metric measures the similarity between image and generated-caption embeddings computed using the MedImageInsight model [14]. One possible explanation is that the retrieval-based generation process encouraged the model to produce captions that remained strongly aligned with the visual content of the image. Since the retrieval stage relied on CLIP-based embeddings, the retrieved examples may have guided the language model toward captions with higher embedding-level similarity to the input image.

6. Conclusions and Future Work

This work presents the joint UFPE–Essex–UNED submission to the ImageCLEF 2026 Caption Prediction subtask. The approach employed a fully training-free RAG pipeline combining Llama 4 Maverick, few-shot prompting, and multimodal retrieval using both CLIP and MedSigLIP embeddings. The experiments demonstrated that expanding the retrieval database and refining the prompt to better align with the dataset's caption style yielded the largest performance gains, while using a medically specialized retrieval encoder provided additional improvements. Using this approach, the system placed 2nd in the Caption Prediction subtask with an overall score of 0.3603, demonstrating that competitive medical image captioning performance can be achieved without fine-tuning large models. Although the top-ranked system achieved higher scores in several text-based and factuality-oriented metrics, our method remained highly competitive in terms of image–text semantic alignment.

In addition to the standard subtask, we also participated in the Caption Prediction Synthetical subtask using a simplified version of the same pipeline. Even without the iterative refinements applied to the standard caption experiments, the 4th place was achieved with an overall score of 0.4625. Notably, the system obtained the highest Similarity score among all submissions in this subtask, highlighting its strong performance on the image–text semantic alignment metric in the synthetic-caption setting.

These results further reinforce the potential of training-free multimodal pipelines for medical image captioning across different captioning settings.

To further validate the contributions of each component in the proposed pipeline, future work will include a more comprehensive ablation study. In particular, we intend to isolate the individual impact of prompt engineering, the size of the retrieval database, and the specialization of the retrieval encoder. For example, additional experiments combining the refined prompt with the reduced 10k-image database will help quantify the isolated effect of prompt refinement, while experiments using MedSigLIP with the initial prompt configuration will enable a more direct evaluation of the retrieval encoder's contribution.

Another important direction for future work is the investigation of domain-specific fine-tuning strategies. Although the training-free approach achieved competitive results, it remains unclear how much additional performance could be gained by fine-tuning either the caption-generation model or the retrieval encoder on the Competition's dataset. In this context, we also plan to evaluate specialized multimodal models, such as MedGemma, to compare smaller domain-adapted models with large general-purpose LLMs.

Additional approaches may also further improve caption quality. One promising direction is the incorporation of iterative feedback mechanisms that assess the similarity between generated captions and image embeddings, enabling the system to refine its outputs based on multimodal consistency. Another interesting possibility is the use of mixture-of-experts strategies combining captions generated by multiple language models to improve robustness and descriptive quality.

Acknowledgments

This work was partly supported by the projects GRESEL-UNED (PID2023-151280OB-C22) funded by MICIU/AEI/ AEI 501100011033 and ANNOTATE (PID2024-156022OB-C31) funded by MICIU/AEI/10.13039/501100011033 and the European Social Fund Plus (ESF+). We also acknowledge partial support from the Brazilian governmental agencies CNPq and CAPES.

7. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT for Sentence Polishing and Rephrasing. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] H. Damm, T. M. G. Pakull, L. Reinartz, B. Bracke, B. Eryılmaz, P. Nath, R. Brüngel, C. S. Schmidt, H. Schäfer, A. Ben Abacha, A. García Seco de Herrera, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2026 – medical concept detection and caption generation with synthetic data extensions, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026. To appear.

[2] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryılmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of ImageCLEF 2026: Multimodal challenges in medicine, science, agritech, and security, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International

Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.

[3] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, Y. Choi, CLIPScore: A reference-free evaluation metric for image captioning, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7514–7528. URL: https://aclanthology.org/2021.emnlp-main.595/. doi:10.18653/v1/2021.emnlp-main.595.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, Neural Information Processing Systems abs/2005.11401 (2020).

[5] H. M. Pereira, P. D. Maranhão, V. de Araújo Xavier, T. I. Ren, A. Duggal, A. G. S. de Herrera, V. Abolghasemi, Evaluating generative AI for medical image captioning: a benchmark study on radiology images, in: X. Yang, W. Hsu (Eds.), Medical Imaging 2026: Imaging Informatics, volume 13930, International Society for Optics and Photonics, SPIE, 2026, p. 139300F. URL: https://doi.org/10.1117/12.3084497. doi:10.1117/12.3084497.

[6] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. S. de Herrera, H. Müller, P. Horn, F. Nensa, C. M. Friedrich, ROCOv2: Radiology objects in context version 2, an updated multimodal image dataset, Scientific Data 11 (2024). doi:10.1038/s41597-024-03496-6.

[7] OpenAI, Gpt-4o system card, 2024. URL: https://arxiv.org/abs/2410.21276. arXiv:2410.21276.

[8] Google DeepMind, Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL: https://arxiv.org/abs/2507.06261. arXiv:2507.06261.

[9] Meta AI, Introducing llama 4: Advancing multimodal intelligence, https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[11] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau, J. Chen, F. Mahvar, L. Yatziv, T. Chen, B. Sterling, S. A. Baby, S. M. Baby, J. Lai, S. Schmidgall, L. Yang, K. Chen, P. Bjornsson, S. Reddy, R. Brush, K. Philbrick, M. Asiedu, I. Mezerreg, H. Hu, H. Yang, R. Tiwari, S. Jansen, P. Singh, Y. Liu, S. Azizi, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Riviere, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos, E. Yvinec, M. Casbon, E. Buchatskaya, J.-B. Alayrac, D. Lepikhin, V. Feinberg, S. Borgeaud, A. Andreev, C. Hardin, R. Dadashi, L. Hussenot, A. Joulin, O. Bachem, Y. Matias, K. Chou, A. Hassidim, K. Goel, C. Farabet, J. Barral, T. Warkentin, J. Shlens, D. Fleet, V. Cotruta, O. Sanseviero, G. Martins, P. Kirk, A. Rao, S. Shetty, D. F. Steiner, C. Kirmizibayrak, R. Pilgrim, D. Golden, L. Yang, Medgemma technical report, 2026. URL: https://arxiv.org/abs/2507.05201. arXiv:2507.05201.

[12] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[13] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[14] N. C. F. Codella, Y. Jin, S. Jain, Y. Gu, H. H. Lee, A. B. Abacha, A. Santamaria-Pang, W. Guyman, N. Sangani, S. Zhang, H. Poon, S. Hyland, S. Bannur, J. Alvarez-Valle, X. Li, J. Garrett, A. McMillan, G. Rajguru, M. Maddi, N. Vijayrania, R. Bhimai, N. Mecklenburg, R. Jain, D. Holstein, N. Gaur, V. Aski, J.-N. Hwang, T. Lin, I. Tarapov, M. Lungren, M. Wei, MedImageInsight: An Open-Source Embedding Model for General Domain Medical Imaging, 2024. URL: https://arxiv.org/abs/2410.

06542. arXiv:2410.06542.

[15] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7881–7892. URL: https://aclanthology.org/2020.acl-main.704/. doi:10.18653/v1/2020.acl-main.704.

[16] Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. A. Folarin, A. Roberts, R. Bendayan, M. P. Richardson, R. Stewart, A. D. Shah, W. K. Wong, Z. Ibrahim, J. T. Teo, R. J. B. Dobson, Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit, Artif Intell Med 117 (2021) 102083.

[17] W.-w. Yim, Y. Fu, A. Ben Abacha, N. Snider, T. Lin, M. Yetisgen, Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation, Scientific Data 10 (2023) 586. URL: https://doi.org/10.1038/s41597-023-02487-3. doi:10.1038/s41597-023-02487-3.

[18] Y. Zha, Y. Yang, R. Li, Z. Hu, AlignScore: Evaluating factual consistency with a unified alignment function, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 11328–11348. URL: https://aclanthology.org/2023.acl-long.634/. doi:10.18653/v1/2023.acl-long.634.

[19] Y. Pakbaz, P. Hoseinpour, F. Olamaeian, N. Nafissi, Innovative technique for managing extreme relapsing bilateral pseudoangiomatous stromal hyperplasia (PASH) in a young woman: A case report highlighting a novel intervention in reconstruction, Int J Surg Case Rep 120 (2024) 109873.

[20] Y. Liang, H. Li, K. Meng, B. Bu, Y. Zhai, M. Li, Rendezvous-assisted endoscopic retrograde pancreatography in groove pancreatitis with a nondilated pancreatic duct, Endoscopy 56 (2024) E435–E436.

[21] H. Damm, T. M. G. Pakull, H. Becker, B. Bracke, B. Eryilmaz, L. Bloch, R. Brüngel, C. S. Schmidt, J. Rückert, O. Pelka, H. Schäfer, A. Idrissi-Yaghir, A. B. Abacha, A. G. S. de Herrera, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2025 – medical concept detection and interpretable caption generation, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[22] clef-caption-evaluation-2026, 2026. URL: https://github.com/taubsity/clef-caption-evaluation-2026, GitHub repository.

[23] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[25] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/.

[26] F. H. V. Cheung, C. C. C. Mak, W. Y. Chu, N. Fan, K. W. Lui, A case of type II mirizzi syndrome treated by simple endoscopic means, J Surg Case Rep 2018 (2018) rjy257.

[27] S. Jawale, Intrapancreatic autologous stem cell therapy for type 1 diabetes - an experimental study, Ann Med Surg (Lond) 85 (2023) 4355–4371.

[28] S. C. Shah, R. Radadiya, A. Patel, S. M. Velamakanni, T. Patel, Percutaneous retrieval of dislodged chemo port catheter with inaccessible tips by a simplified technique, Cureus 14 (2022) e21692.

[29] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: ICCV, 2023, pp. 11941 – 11952. doi:10.1109/iccv51070.2023.01100.